DOCUMENT RESUME

ED 306 241 TM 013 005

AUTHOR Franklin, Jennifer; Theall, Michael
TITLE Who Reads Ratings: Knowledge, Attitude, and Practice of Users of Student Ratings of Instruction.
PUB DATE Mar 89
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Faculty; College Students; Course Evaluation; *Evaluation Utilization; Evaluators; Formative Evaluation; Higher Education; Knowledge Level; School Surveys; Student Attitudes; *Student Evaluation of Teacher Performance; Summative Evaluation; *Teacher Evaluation

ABSTRACT

What users of student ratings of instruction know about ratings and what they need to know to make valid and reliable use of ratings data were studied. An item bank of 153 multiple-choice items was developed from sources of information about student ratings. The full item bank had five subsets: (1) knowledge of ratings concepts; (2) knowledge of quantitative issues for interpreting/applying ratings; (3) simulated practice exercises; (4) attitudes toward ratings issues; and (5) demographic variables associated with rating users. Items were validated by a panel of 24 experts to develop the shorter questionnaires mailed to faculty at three institutions and members of a professional group for a total of 779 respondents. Participation was 15% for one institution and 20% for each of the other groups, except that the participation of the 23 experts in the field of ratings use was 100%. The magnitude of the difference between the expert scores and the scores of ratings users confirmed that users may not know all that they need about using student ratings. Those with a negative attitude were likely to use student ratings less often and less well. Positive attitudes were associated with higher knowledge scores. Many users of ratings appear insufficiently aware of the issues currently set forth in ratings literature to be able to make decisions for summative or formative purposes. Five data tables are included. Appendix A is the 69-item ratings survey, and Appendix B lists the names and academic affiliations of the experts from the first validation. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306241

Who Reads Ratings: Knowledge, Attitude, and Practice of Users of Student Ratings of Instruction

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JENNIFER FRANKLIN
MICHAEL THEALL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Jennifer Franklin

and

Michael Theall

Office of Instructional Development and Evaluation
Northeastern University
Boston, Massachusetts 02115

Presented at the Annual Meeting of the American Education Research Association.

San Francisco, 1989

Please do not cite or reproduce without permission.

TM013005

WHO READS RATINGS: KNOWLEDGE, ATTITUDES, AND PRACTICE OF USERS OF STUDENT RATINGS OF INSTRUCTION

1. Introduction

Reports of results are a critical link between student raters and evaluators. Nonetheless, studies probing the ways users interpret results or the roles of report variables such as content and format are nearly absent from the literature. The design of reports has received almost no attention, nor have the qualifications of the evaluators who read them. As a result, there is insufficient evidence to conclude that the results are understood and applied by users with at least the validity and reliability of the instruments that obtained them.

Students provide evaluative data when they are asked to respond to items such as "rate this instructor compared with others". They provide descriptive data when they describe the behavior of the teacher by responding to items such as "the instructor presented information at a rate I could follow". It would be easy to mistake students for the evaluators in the student rating system because students provide their own value judgements, but the real evaluator is actually the one who uses student data to make judgements of merit, worth, or effectiveness.

Anyone who interprets student data for the purpose of applying it to some other process becomes an evaluator. Administrators evaluate faculty teaching performance for personnel decision-making. Faculty evaluate their own performance to provide evidence in their merit review portfolios for and for improving their own teaching and course design skills. In the latter case, faculty can be both the evaluator and the evaluated. Specialists such as instructional developers and teaching improvement consultants are evaluators when they interpret ratings in order to help faculty assess teaching skills, evaluate instructional materials, and/or develop courses.

Evaluators can collect and report the data they use or can be clients of ratings systems which collect and report data. Results provided by evaluation data processing services are generally transmitted to clients in printed reports or in personal interaction with teaching improvement consultants. Specialists can bring their training and experience to bear on the interpretation of results in ways that will be helpful to clients. Without assistance, clients must bring skill to the process of decoding statistical results into meaningful information.

All components of instructional systems need periodic evaluation. Ratings systems should not be exempt. There are immediate reasons to be concerned about how well ratings are working. There is abundant anecdotal evidence that many faculty continue to challenge the validity of student ratings as a meaningful measure of teaching performance or course design. They are cynical about students as raters and about administrative uses of ratings data. Some believe that using ratings demeans the teaching mission by treating it as a consumer product. Although persistent negative attitudes could be explained as the result of misinformation and

ignorance about evaluation, lack of reward for effective teaching, or risk of exposure for poorly rated faculty, it is also possible that real abuses of ratings data are generating hostility and cynicism.

2. Background

If the reporting component of the evaluation cycle is an integral part of the process, research literature should be a source of information about reporting results. But no blueprint or instructions for communicating with users will be found there, because four decades of research literature have focused almost entirely on one aspect of student ratings: measurement.

Handbooks are probably the most important source of information about ratings for practitioners. Handbooks are extended works offering general theories and recommendations for practice about teaching or the evaluation of teaching. They synthesize the research literature for non-research oriented readers and often provide overviews of the general processes of merit review and teaching improvement. Typically, they describe the characteristics of good ratings instruments and proper procedures for administering them. They provide lists of important research results that may signal sources of bias or systematic differences in student ratings, e.g., teacher, course, and student variables.

How ratings issues are treated in handbooks is usually based on the context in which ratings will be used: merit review (Braskamp et al, 1984; Centra, 1980; Doyle, 1983; Miller, 1988; Millman (Ed), 1981; Seldin, 1984) or college teaching techniques (Bergquist & Phillips, 1975; Furmann and Grasha, 1983; Guillette, 1982; McKeachie, 1978; Lowman, 1984; Eble, 1985). Often the orientation of the author seems to determine the general approach to explaining how to use ratings. Handbooks for the appraisal of faculty performance are particularly concerned with implications of the research literature for fair practice in merit review. For example, the dictum that ratings results should only be used along with other sources of information about teaching performance is typical (Braskamp et al, 1984; Centra, 1980; Doyle, 1983;Seldin, 1984). Likewise, considerable attention is usually paid to the issue of sample size and composition for comparing ratings. Alternatively, handbooks on teaching technique aim to persuade faculty that student feedback is a valuable source of information for improving teaching and course design skills and therefore tend take a more qualitative approach. (Guilette 1982; McKeachie 1978).

Although many handbooks provide general recommendations concerning how results should be reported, discussions tend to center on the statistical models needed for the job (e.g. appropriately normed comparison statistics for personnel decision-making and descriptive statistics for teaching improvement). Moreover, these discussions are brief, usually a paragraph or two. Information about graphic formats, explanatory content, and other aspects of reports are not offered in enough detail to provide standards or procedures to guide the development of reports. Only one handbook (Doyle, 1983) has a full chapter on collecting and reporting data. While every one of the 'classic' handbooks has at least one example of an exemplary ratings instrument, only one (Braskamp, et al 1984) has an illustrated example of a report of results although there is no discussion of what is important to notice about the report.

Doyle (1983) discusses the problem of representing quantitative information. He offers possibilities ranging from tabular arrays of numbers or graphic depictions such as bar charts, to result-contingent, individualized messages concerning interpretation and teaching improvement issues. The last possibility has become more feasible with the advent of low cost computing and increasingly sophisticated software. Although Doyle offers the most comprehensive source of ideas for the production of reports for various purposes, a theory of how, what, and why to communicate to recipients is still missing.

There are other sources for information about ratings and reports of results. Nationally disseminated ratings services such as Kansas State's IDEA and Educational Testing Service's SIR (Student Instructional Report) provide good support for users. Reports from these systems always include adjunct text materials explaining how to read the tables of numbers along with cautions and advice concerning sources of bias in ratings. A chief strength of these systems is their long-standing and comprehensive norms for faculty in various content areas.

Local handouts explaining relevant statistical concepts are frequently provided with the numeric tables of results, especially when professional evaluators, measurement specialists, or instructional systems developers are involved. Institutions with large in-house instructional development shops can produce more elaborate materials. A publication from Syracuse University, "A Guide to Evaluating Teaching for Promotion and Tenure", (Centra et al, 1988) is a particularly good example of detailed information, including what may be the only illustrated, suggested examples of reports formats for specific purposes such as communicating results to administrators. Kansas State University's "Idea Papers" combine the power of a nationally distributed ratings system with the resources of local faculty/instructional development shop to provide a very valuable source of information about teaching and ratings.

Training workshops and seminars such as those sponsored by Kansas State, or private consultation services such as Aleamoni and Arreola's Comprehensive Date Evaluation Services (CODES), may offer suggestions on reporting results along with general information about ratings and their uses. However, these events vary in detail and direction since the major foci of these workshops are the larger issues of the role of ratings in merit review at the departmental or institutional level and the development of evaluation systems.

Certain professional organization publications regularly offer information about ratings. "Instructional Evaluation", the quarterly publication of the American Education Research Association Faculty Evaluation and Development Special Interest Group, has become an important forum for discussing ratings research and practices. "To Improve the Academy" a joint publication of the Professional Organizational Development Network in Higher Education (POD) and the National Council for Staff, Program and Organizational Development is a frequent source as well.

Turning to the other side of the process of communicating results, if little has been written about reports of results, less has been written about those who use them. Levinson-Rose and Menges (1981) in their major review of interventions to improve college teaching concluded that feedback with consultation may improve instruction more than written feedback alone and that feedback has the most positive effect on those who rate themselves higher than they are rated by their students. The studies reviewed provide little concrete information about the printed reports used. Aleamoni, however, does provide a description of a "computerized" report and refers to a manual of instructions provided to subjects in a study comparing feedback of results

and feedback of results with consultation. (Aleamoni, 1978). However, how well manuals work, is unknown, since they are rarely, if ever, systematically evaluated as instructional material.

Such findings are encouraging in that student ratings can be used to improve teaching, but the studies reviewed to find these conclusions typically do not consider the effects of report variables. Printed reports of results do not have to be exclusively quantitative; e.g. expert systems software offers the possiblity of useful, inividualized explanatory messages. It does not seem likely that all reports are any more equivalent in their usefulness in improving instruction than are all student rating questionnaires. Yet, whether a questionnaire contains items describing teaching behaviors along with global, value-based items (e.g. "rate this instructor compare to others") certainly will determine the appropriateness of the data it produces for teaching improvement purposes. In other words, not that consultants do not help or shouldn't be used, but how can the printed reporting system be improved ?

Users have to be seen in the matrix of problems that surround the use of student ratings. Collecting student ratings is a time-consuming process conducted for practical ends: to inform decisions or actions. The difficulty of reaching consensus on questionnaire items can derail or delay development processes indefinitely. Just getting the most basic system up and running can become an ordeal far beyond the expectations of its developers. The process of evaluating the effectiveness of an evaluation system may raise issues that have only recently been retired. In many settings, mandatory ratings for promotion and tenure can adversely affect the use of ratings for teaching improvement. Even the fact that ratings are in use may tend to create a vested interest in the assumption that ratings users have sufficient skill and knowledge to correctly interpret and apply data: that is, "if you don't know it's broke, you don't have to fix it."

Sorting out which aspects of these problems apply in a given situation is probably necessary in any case, but understanding if there are individual or group differences in ratings users that affect the quality of practice is also important. Knowing the goals of users and the state of their knowledge and skills is essential to make the ratings process work for them.

3. This study

We began with the view that a student ratings system is an aggregation of processes intended to collect opinions about instruction. Among its processes are those which: assure valid measurement; transform collected data into coherent information; and transmit information to users. In addition to these processes, a means for assessing the performance of the components is necessary to understand how well the system works as well as how to make it work better. However, no standards for the performance of ratings systems have been described in the literature of student ratings. Worse, using the literature to justify standards for assessing these systems is difficult in the absence of a ratings theory grounded in more general theory of instruction and teaching improvement. Without such a theory one cannot test the effectiveness of a system.

We held two assumptions: 1) the best ratings systems should operate in ways most consistent with the evidence presented in the research and practice literature and (2) users most

likely to use data appropriately would be those whose knowledge and beliefs about ratings most closely resembled that of the reseachers and practitioners who created the best literature.

The notion "appropriate use of ratings", has not been operationally defined in the literature for practitioners but is left to inference instead. Ultimately, appropriate use requires that when any information derived from ratings data is used to inform formative or summative evaluation, it is valid and reliable in its representation of some state of affairs in the domain of teaching performance. This is only a necessary condition for good evaluation practice, not a sufficient one nor the only one.

The lack of operational definitions for good practice should be a signal that evaluation of the systematic approach to student ratings is needed. Such an evaluation will necessarily include assessment of users and channels for communication if it is to be effective. However, personnel decision making and teaching improvement consultation are by nature confidential processes and therefore not readily open to investigation. Overt scrutiny would likely prejudice these processes in any case. Moreover, many institutions and departments have not yet adopted the kind of standardized, documented merit review processses that would most readily withstand investigation. So observing the users at work under naturalistic conditions may be very difficult.

In the meantime, we can assume that appropriate use depends on in part on the "validity and reliability" of users of ratings information. Looking at basic levels of knowledge and attitudes in users is one strategy for assessing it. Recognition of the "facts" of evaluation should be evident in those who are able to use ratings appropriately. Lack of recognition would imply that such users are at greater risk of perpetrating bad practice. For example, users should be aware of the student, teacher, and course variables that have consistently or even intermittently been shown to related to ratings outcomes. They should also be aware of the characteristics of properly constructed ratings instruments and the condition necessary to obtain valid, reliable results. Most of all, since ratings are typically expressed quantitatively, users should be able to understand numerical information and to avoid making decisions based on chance differences.

Generally positive attitudes towards ratings should also be associated with better practice, since users will be in less danger of the consequences of bad practice. It seems likely they will have generally positive attitudes toward the honesty of students as raters and the potential benefits ratings can confer on the quality of instruction. Having had contact with a skilled interpreter of ratings should also have an effect on the attitude and knowledge of users. Moreover, the more knowledge a user has about how to use ratings, the better the user's attitude should be. Thus, it follows that the practice of using ratings will be more appropriate the more nearly users meet these qualifications.

Because the impetus for this study was essentially pragmatic, that is the need for information on which to base the design of reports of results, the goal of this study was to help answer two questions.

1. What do users of ratings currently know about ratings compared to those who created the literature?

and, more pragmatically,

2. What do users need to know in order make valid and reliable use of ratings data?

4. Method

Instrument

A collection of 153 multiple choice items was developed from a comprehensive review of sources of information about student ratings. Concepts were selected for face validity and based on: direct paraphrases of sources; generalizations relating to important aspects of practice directly recommended by or clearly to be inferred from sources; general quantitative knowledge needed to understand typical five point scaled ratings results; and attitudes associated with the use of ratings. Several additional items were constructed or rewritten, totaling an item-bank of 183 items. (See Appendix A for samples of items.)

The full item bank consisted of 5 subsets.

1. knowledge of substantive ratings concepts including: validity and reliability issues, 16 items; knowledge of adminstrative and procedural issues, 15 items; knowledge of student variables affecting ratings outcomes, 11 items; knowledge of teacher or course variables affecting ratings outcomes, 11 items; knowledge of qualitative issues essential for interpreting/applying data to decision-making, 10 items
2. knowledge of quantitative issues essential for interpreting/applying data including specific statistical models and sampling issues, 30 items
3. simululated practice exercises: using results for promotion and tenure, 21 items; for teaching improvement, 32 items.
4. attitudes toward ratings issues, 23 items
5. demographic variables associated with ratings users, 9 items.

Knowledge Items:

Each item was constructed as statement with which a knowledgable user of ratings would either agree, disagree, or indicate uncertainty/no opinion. Each statement was presented as a matter of opinion by using a "strongly agree" to "strongly disagree" response scale. The intention was primarily to make the experience of responding as neutral as possible, avoiding the appearance of a test by soliciting 'opinions' rather than 'answers'. In many cases, items were direct or paraphrased versions of statements made in handbooks or articles providing instructions for the use of ratings. It is important to note that the content of items was relatively general and intended to have high face validity. Esoteric concepts were avoided whenever possible. For example, the fact that the student ratings are positively correlated with student achievement is a cornerstone of validity arguments for ratings and is well established. (Cohen, 1981). Therefore a knowledgable respondent would have to disagree with the statement:

> "There is no predictable relationship between student ratings and student achievement."

Because lack of respondent anonymity can bias ratings, a knowledgable respondent should agree with (Seldin, 1984):

> "Student ratings used for personnel purposes should not be returned to the instructor until after final grades have been submitted."

Attitude items:

Attitude items were constructed to identify respondents' general attitudes towards ratings and their use in the respondents' own experiences.

For example, a respondent with positive attitude toward ratings would likely disagree with the statement:

> "I feel my career has been harmed to some degree by student ratings I have received."

and agree with:

> "I trust my students to give sincere and honest responses to student ratings questionnaires."

Demographic items:

These items were intended to identify characteristics of faculty such as rank, years of teaching experience, general participation in various ratings processes, and participation in promotion and tenure or teaching improvement applications of ratings for themselves and others.

Practice items:

This subset was intended to probe knowledge of standards for practice and the application of knowledge of quantitative issues to decision-making. The items used the same "Strongly agree to Strongly disagree, or uncertain" scale as the knowledge items. The preferred answer was "uncertain" because insufficient data were deliberately presented.

A simulated report of results (for a group of 5 faculty with 2 to 5 courses for each) was presented along with a set of statements characterizing the results. This report offered the most commonly advocated statistics for reports of ratings in an array appropriate for a committee or administrator. The respondent was asked to assume s/he was a member of a departmental promotion and tenure committee. The task was to indicate agreement with statements representing judgements of merit or worth made by fellow committee members. The simulated

report of results contained means and standard deviations for the institution and the department as well as T scores, and percentile ranks in one of five groups (top 10%=HI, next 20%=HI MID, middle 40%=MID, next 20%=LO MID, bottom 10%=LO). (See Appendix A.)

For example, one item in this section asks if Professor Fahey is a better teacher than Professor Cohen, based on the simulated ratings report. The principle that the proportion of respondents in each class is as important as the number of respondents is well documented (for example, Seldin, 1983, Centra, 1980, Braskamp et al, 1984) and was violated in this data. Therefore, If Smith's report shows five small classes (less than 20 students enrolled) with a response rate less than fifty percent in each case, one should draw no conclusions based on those data. Thus, in this exercise, one should indicate for every judgement of merit or worth involving that instructor that there is insufficient evidence on which to base a judgement. The only valid response has to be "uncertain".

In part 2 of this subset, diagnostic judgements were solicited for the purpose of teaching improvement consultation. These items were not used in subsequent surveys although the data from this exercise were used to develop open-ended items concerning teaching improvement consultation for the survey sent to the professional organization. No further analysis for these items is presented in this report.

5. Construct Validation

The items were submitted to a panel of 24 expert consultants (see Appendix B) selected for their participation in the development of the literature on which the items were based. Consultants were selected for their contributions to one or more of three areas: research in ratings measurement issues, teaching improvement consultation using ratings, and development and/or administration of large scale commercial or institutional evaluation services using ratings. In part, citation counts and quoting circles helped confirm the selections.

The consultation group was asked to respond to each item with what they believed to be the best response. Open-ended comments were also solicited using a matrix with a brief indication of the author's preferred answers and spaces for comments next to each item. Consultants were asked not to look at matrix materials and answers until their own answers had been recorded.

In general, expert agreement was obtained on nearly all items (using majority opinion as a measure). Majority opinion was in the predicted direction for all but 4 items of the original 131 knowledge and practice items. One set of consultant responses was set aside because it were incomplete. Because expert opinion was used to validate items, those which did not produce at least 74% agreement were discarded from subsequent surveys or, in a few cases, salvaged by rewriting.

Some items simply performed poorly because of defects in construction; others revealed real disagreements among experts. Two particular kinds of disagreement appeared: disputes concerning the relative importance and meaning of research findings, and disagreement about matters of practice.

Examples of disagreement over research findings were prone to appear in matters such as the effects of class size, course level, and certain teacher characteristics. Disagreements over practice were most frequent in the scale that used a simulation strategy to force interpretations of ratings data for promotion and tenure decision-making. Based on the open-ended comments provided by the experts, there was disagreement about the consequences of using "less than perfect" samples. In the expert group, there appear to have been strict constructionists, pragmatists, and possibly those who may have simply erred. Lack of expert consensus on these items is less disturbing than if ratings users with far less knowledge and skill than the expert group are using substantially flawed data.

The differences among expert raters fell to some extent along the lines of research versus practice. However, the "researcher" types were more prone to take exception to particular constructions within items than were those who operate ratings systems and those who provide teaching consultatin using ratings on a regular basis. The difficulty of writing items that make sense to expert respondents and still can be parsed by novice respondents is particularly challenging.

In addition to tabulating percent agreement as a measure of validity, inter-item correlations for expert responses were analyzed producing the following scale alphas.

Scale 1	knowlege of validity and reliability issues	a = .76
Scale 2	knowlege of quantitative concepts for ratings	a = .72
Scale 3	knowledge of standards for practice, i.e. simulated practice item	a=.70
Scale 4	attitudes towards ratings (.60 without item 92)	a=.53

Several items in Scales 1, 2, and 4 had 0.00 variance. The attitude scale had the smallest number of items and hence probably achieved a lower alpha. Alphas for scales 1, 2 and 4 were considered good, given the relative heterogeniety of items within subsets and small numbers of items. The fact that scale 3 items had relatively high internal consistency (alpha = .70) suggests that a closer look at these results is needed before concluding that validation of these items was thoroughly accomplished.

Survey instrument development:

Based on results of the expert responses and comments, shorter questionnaire forms (75, 69 and 55 items, respectively) were developed. In some cases, items were revised to make them read more clearly. A few new items were constructed. Each survey contained some items in common with the other surveys as well as unique items. The 75 and 69 item forms shared 50 items with the original expert validated set and included several major rewrites of original items. The 55 item form shared 34 items with the original set, 8 with the 75 and 69 item forms, and 13 new items were oriented specifically toward the role of ratings in teaching improvement consultation. No attempt was made to weight items according to the relative importance of their underlying constructs.

6. Method

Data Collection

The revised questionnaire forms were mailed to faculty at large in three institutions: a large private university, "School A"; a large multi-campus state university, "School B"; and a large community college system, "School C". A shorter, 55 item form was sent to 420 members of a professional association of faculty, adminstrators, and instructional technologists interested in issues of effective post-secondary teaching and faculty development. TABLE 1 below shows the number of respondents, percent response to mailing, and the length of the form in each group.

TABLE 1. SURVEY SAMPLES

SURVEY GROUP	respondents	partic.%	n items on form
Expert Consultants	23	100%	153
School A	193	15%	75
School B	245	20%	69
School C	232	20%	69
Prof.Org	85	20%	55
total respondents =	779		

Participation rates ranged from approximately 15% to nearly 20%, considered a good response rate for a mail survey with no opportunity for follow-up. Respondents included all ranks of faculty ranging from full professor to teaching assistant. The largest single group however was "assistant professor". (See Table 2, opposite)

Because the commitment of institutional resources for the improvement of teaching may be a factor in attitudes toward ratings, survey groups were selected to represent levels of support as well as institution type. The institutional climates of the three faculty sample groups varied in history of ratings use and support for teaching/faculty development:

School A: mandatory evaluation of undergraduate courses for publication in a student rating course selection catalog, uregulated but widespread use of dpartmentally developed ratings for promotion and tenure; relatively little institutional support for teaching improvement

School B: widespread and mandatory use of ratings along with a relatively well supported, long-standing agency for teaching improvement

TABLE 2.

DEMOGRAPHIC VARIABLES FOR SURVEY GROUPS (%)

	GROUPS					
CATEGORY	Experts n=23	School A n=193	School B n=215	School C n=232	Prof-org. n=85	All Groups n=779
RANK						
full professor	64	22	18	35	30	30
associate professor	27	27	31	30	35	30
assistant professor	5	3	24	30	15	25
instructor	0	3	18	0	0	8
lecturer (and part-time)	5	11	4	0	10	6
teaching assistant	0	6	5	5	10	2
YEARS TEACHING						
less than 1 year	0	5	5	0	-	3
1 to 2 years	0	7	11	5	-	5
more than 2 less than 8	9	28	16	20	-	22
more than 8 less than 12	17	20	10	20	-	18
more than 12 less than 20	35	16	18	10	-	28
more than 20	40	25	42	45	-	27
ROLE						
teaching only	30	76	70	63	56	62
teaching, some admin.	22	16	18	21	39	18
admin, some teaching	13	4	5	11	0	6
administration only	0	1	0	5	0	9
instruc/faculty dev, some teaching	26	2	4	5	0	9
instruc/faculty dev. only	9	1	2	0	0	3
PERSONNEL DECISION-MAKING						
yes, currently	36	34	24	60	48	23
yes, not currently	46	20	21	10	14	26
no, never	18	46	55	30	38	51
RECEIVED TEACHING CONSULTATION USING RATINGS						
yes, 2 or more times	38	7	21	0	0	17
yes, but only once	0	5	16	5	5	10
no, but I'd like to try it	62	65	53	80	65	59
no, and I don't want to try it	0	23	11	16	30	15

School C: strong reputation for faculty development and the support of effective teaching but with little systematic use of student ratings or consultation for teaching improvement.

Prof.Org. The professional and organizational development mission of the group meant that these respondents would be, to a great extent, self-selected for interest in or experience with student ratings regardless of their institution's level of support for teaching.

The questionnaire and machine-scorable answer sheet were accompanied by a letter explaining the general goals of the project, why respondent input was important, and that their institution would receive a general summary showing their responses compared with the expert, professional, and faculty groups. Many respondents offered unsolicited written comments ranging from comments agreeing with the need for such a study to out-and-out hostilty toward "educationists" who would perpetrate ratings in the first place.

7. Analysis

We are reporting essentially descriptive information, although some statistical tests (SPSSX ONEWAY) analysis of variance were applied to apriori assumptions about the direction and magnitude of our results.[1] In most cases, statistical tests for comparisons of means were based on the knowledge and attitude items common to all questionnaire forms. Differences in the attitudes and knowledge were examined with respect to several demographic items

Associations between knowledge items and attitude items were the subject of correlation (SPSSX CORRELATION) analysis.

Scoring:

Although the knowledge scales were presented to respondents as multiple choice items (agree to disagree with an uncertain option), analysis of results was based on recoding these items as true/false with correct and incorrect answers. "Tend to agree" and "strongly agree" options and "tend to disagree" and "strongly disagree" options were collapsed into agree and disagree.

Following tabulation of frequencies for all items, results for knowledge items were expressed as percent correct of total items. Attitude items were expressed as means on a 6-point scale, and frequencies as percent response for demographic items. Items were scaled in the

[1] Because the number of respondents in the expert group was very small (n = 23) and variance was notably smaller for all scores in this group compared with all other groups, Bartlett's Box test for homogeneity of variance was performed in order to detect signi ficant differences in variance among the survey groups. Although variance was significantly different for almost all comparisons involving the expert group, the smallest variance was always associated with the smallest group. According to Hopkins and Glass, 1984, in such a case the Type I error rate is actually lower than the obtained Anova value.

direction of the majority expert preference. In general, only items on which experts reached or exceeded 74% agreement were included in susequent surveys. High point value was assigned according to the direction of responses of experts when the item involved a positive attitude towards ratings. "Attitude" items which did not involve value issues were scored separately and treated as demographic items. Demographic items (subset 5) were recorded as frequencies and analyzed as percent by option.

Because of the many items and multiple questionnaire forms involved, two composite scales (COMMON 1 and COMMON 2) were devised based on the knowledge probing items within subsets 1,2,and 3 appearing on every form. The number of items in common for the expert and faculty groups (COMMON 1, n=25) was considerably larger than that for the expert,faculty groups, and the professional organization together. (COMMON 2, n=17). Attitude scales (ATT 1 and ATT2) were also created for the attitudes items on the same basis.

8. Results

Knowledge items

Table 3 shows results as mean percent correct with 95 percent confidence intervals for each subset for each group . Because the number of items on each group's questionnaire varies on each scale, this table is intended strictly to represent the results for each group with no comparisons among groups intended. Note, for each score, 95% confidence intervals are relatively narrow.

TABLE 3. KNOWLEDGE SUBSET MEAN SCORES (PERCENT CORRECT)

	GROUPS									
ITEM SUBSET	EXPERTS		SCHOOL A		SCHOOL B		SCHOOL C		PROF-ORG	
	x̄	95%c.i.	x̄	95%c.i.	x̄	95%c.i.	x̄	95%c.i.	x̄	95%c.i.
SUBSET 1:	84	(81-88)	50	(48-52)	52	(51-54)	53	(52-55)	61	(58-63)
SUBSET 2:	83	(77-90)	51	(48-53)	51	(49-53)	44	(41-46)	50	(49-51)
SUBSET 3	69	(63-76)	61	(56-66)	61	(57-65)	62	(58-66)	62	(60-64)

SUBSET 1: knowledge of validity and reliability issues:
SUBSET 2: knowledge of quantitative issues
SUBSET 3: knowledge of standards for practice

Table 4 compares means, standard deviations, and the value of F for each of the the composite scales. COMMON 1, COMMON 2, ATT 1, AND ATT 2. It also shows COMMON 2 by "participation in promotion and tenure decisions using ratings", COMMON 2 by "received teaching improvement consultation using ratings" and COMMON 2 by respondent's institutional function. For the differences between groups, mean percent correct for each group of survey respondents was significantly lower than scores of the expert group. In general, distances between school groups' scores were small, while the professional organization's scores

TABLE 4. DIFFERENCES IN MEAN SCALE SCORES

SCALE BY GROUP Group	$\bar{x}$	s.d.	min	max	F
COMMON 1					
Experts	81.6	8.9	63.5	92.3	
School A	49.7	13.4	3.8	84.6	
School B	52.7	12.9	19.2	80.7	
School C	50.1	12.1	3.8	76.9	F = 45.93, p < .000
COMMON 2					
Experts	82.4	10.8	58.8	94.1	
School A	49.6	16.0	5.88	94.1	
School B	49.0	16.4	5.88	88.2	
School C	47.2	14.4	5.88	82.3	
Prof.Org.	63.7	14.6	35.3	94.1	F = 43.54, p < .000
ATT 1					
Experts	4.62	.46	3.50	5.50	
School A	3.64	.84	.10	5.00	
SchooL B	3.86	.55	1.00	4.90	
School C	4.04	.50	2.20	5.50	F = 24.61, p < .000
ATT 2					
Experts	4.60	.35	3.83	5.17	
School A	3.77	.79	.17	5.17	
School B	3.82	.66	.83	5.17	
School C	3.98	.55	2.00	5.33	
Prof. Org./	4.20	.50	3.00	5.17	F = 15.38, p < .000
RECEIVED ASSISTANCE INTERPRETING RATINGS FOR TEACHING IMPROVEMENT BY COMMON 2:					
Yes	56.70	16.1	5.88	94.11	
No	50.70	18.3	5.88	94.11	F = 13.91, p < .000
PARTICIPATED IN MERIT REVIEW OF OTHERS USING RATINGS BY COMMON 2					
Yes, regularly	54.4	17.9	5.88	94.11	
Yes, but not currently	54.2	18.1	5.88	94.11	
No, never	50.2	15.8	5.88	94.11	F = 4.20, p < .02
INSTITUTIONAL FUNCTION BY COMMON 2					
instruc. dev. /eval.	61.8	19.0	29.4	94.11	
instruc.dev./eval/teach	64.6	16.7	29.4	94.11	
administration only	50.9	19.1	29.4	88.24	
admin. some teaching	57.7	17.1	17.6	94.11	
teach , some admin	51.5	18.1	5.88	94.11	
teaching only	49.7	15.1	5.88	94.11	F = 10.04, p < .000

Student-Newman-Keuls post hoc comparisons: p < .05

COMMON 1: EXPERT > COLLEGE B > ind
COMMON 2: EXPER > PROF ORG > COLLEGE B > ind
ATT 1: EXPERT > SCHOOL C > SCHOOL B > SCHOOL A
ATT 2: EXPERTS > PROF ORG > COLLEGE C > COLLEGE B > COLLEGE A

were significantly higher than the school groups , but still considerably lower than the experts scores . These results generally held true for each comparison of knowledge subsets 1 and 2 among all groups. Moreover, all school group scores were lower than the professional organization's scores. The order of differences shown at the bottom of the table is significant at the .05 level. (Student-Newman-Keuls post hoc comparisons)

No significant association between scores on SCR1 or SCR2 and the knowledge of standards for practice subset (SCR3) was found. In fact, the association of SCR3 with the COMMON 1 and COMMON 2 was negligible (r=.15, p<.000 and *n.s.*,respectively). Although the practice items on these scales produced 75% agreement among experts,

For all subjects combined, correlation between the knowledge of validity and relibility subset(SCR1) and the knowledge of quantitative issues (SCR2) was significant (r=.36, p<.000).

Analysis of variance across groups by demographic variables showed no significant differences in COMMON 1 or COMMON 2 based on instructor rank , years teaching, or participation in creating, adapting, or selecting rating forms.

Demographic variables including institutional function, participation in promotion and tenure decision-maing using ratings, and receiving teaching improvement consultation using ratings all produced significant differences in COMMON1 and COMMON 2 means. Those who reported instructional development roles scored higher on COMMON 1 and COMMON 2 than those who reported adminstrative combined with teaching roles. These respondents scored higher than those who only teach followed by those who have only administrative duties. ($F_{(631,5)}$=3.38, p<.005 and $F_{(573,5)}$ = 10.04 , p < .000). The larger sample of instuctional developers /evaluators for COMMON 2 may account for COMMON 2 having a higher mean.

Significant differences in overall percent of items correct for COMMON 2, were obtained based on whether the respondent had received assistance in interpreting ratings for the purpose of teaching improvement . Those who reported receiving assistance ($\bar{x}$=56.7) scored significantly higher on overall number of items correct than those who hadn't ($\bar{x}$=50.7). These results were more dramatic when those who had received assistance more than once ($\bar{x}$=54.5) were compared with those who had never received assistance and would not like to do so ($\bar{x}$=43.9).

Whether the respondent had participated in promotion and tenure decisions produced a small magnitude, but significant difference in COMMON 2: those who had never participated ($\bar{x}$= 50.2 , n=289), vs. those who had previously done so or do so regularly ($\bar{x}$=54.4, n=297).

Attitude items

Table 5 shows attitude item means for each group for the unique set of items. (Scale means for attitude items are shown on Table 4.) The magnitude of differences in attitudes among survey groups is similar to the knowlege dimension , for ATT1 and ATT2 . The positive relationship between attitude and knowlege was also confirmed in many survey groups. For the non-expert survey groups, positive associations between overall score COMMON 2 and ATT 2 were obtained (r=.38, p<.000). For the expert group, lack of variance limited the opportunity to demonstrate the same relationship.

TABLE 5: ATTITUDE SUBSET MEAN SCORES

ITEM SUBSET	EXPERTS x̄ 95%c.i.	SCHOOL A x̄ 95%c.i.	SCHOOL B x̄ 95%c.i.	SCHOOL C x̄ 95%c.i.	PROF-ORG x̄ 95%c.i.
SUBSET 4	4.29 (4.10-4.47)	2.96 (2.87-3.05)	3.17 (3.11-3.23)	3.28 (3.23-3.33)	3.24 (4.13-4.35)

SUBSET 4: attitudes towards ratings (6 , most positive - 1, most negative)

A positive association between the quantitative knowledge subscale SCR 2 and ATT 2 scale ($r=.20, p<.000$) was found and held true for every subgroup but the community college group, SCHOOL C. Many interesting associations between specific attitude items and knowledge scores were also obtained for each group. For the expert group, the specific attitude item: "good teaching is more an innate talent than an acquired skill" showed a relatively strong positive association ($r=.44, p<.01$). For experts, response to "there are few rewards for quality of teaching performance other than personal satisfaction" was inversely associated with high knowledge scores ($r=-.32$, $p<.000$), that is, higher knowledge was associated with the attitude that there are rewards beyond personal satisfaction. For the School A group, this item produced the opposite result ($r=.32, p.<000$), that is, low knowlege was associated with negative attitude.

For School A, there were positive associations between SCR1 scores and "I trust my students to give honest and sincere responses" and ($r=.42, p<.000$); "students have a right/responsibility to rate courses/instructors" ($r=.35 p<.000$); and disagreement with "I feel my career has been harmed to some degree by student ratings I have received. ($r=.35, p<.000$). For School B, only "I trust my students...." was associated with SCR1 scores ($r=.28, p<.000$).

For School C, no associations between specific attitude items and SCR1 were obtained. However, a positive association was found for the quantitative subset score (SCR2) and the item "I know enough about statistics to interpret the results of student ratings without assistance." ($r=.21, p<.000$). There were positive associations between the overall knowledge score (COMMON 1) and the following items: "I trust my students..." ($r=.40, p<.000$); "Student ratings are not a very useful way to measure teaching performance in my particular discipline" ($r=.32$, $p<.000$); "I have been able to improve my teaching based on information I obtained from student ratings" ($r=.25, p<.01$); and disagreement with the statement "I don't really know how to use the results of student ratings to improve my teaching skills" ($r=.19, p<.04$).

For the professional organization group, a positive association between SCR2 (the quantitative scale) and attitude items appeared for "I know enough about statistics to interpret my results without assistance" ($r=.28, p<.005$). An inverse relationship was observed for "I am not aware of any explicit procedures or policies regarding the use of student ratings in personnel decisions in my institution." ($r=-.22, p<.02$). A positive association was found between the overall score, COMMON 2, and disagreement with the item, "I don't really know how to use the results of student ratings to help others improve their teaching skills." was found. ($r=.17, p<.05$).

No significant association between attitude (ATT 2) and institutional function were found. This is notable because knowledge (COMMON 2) and attitude (ATT 2) are consistently

associated with each other for all groups, while institutional function produced a significant difference in COMMON 2 means.

9. Discussion

A major goal of our study was to look for group differences among respondents according to affiliation, demographic variables, and attitude. These differences might help us identify characteristics of users that "put them at risk" when interpreting ratings results.

The magnitude of the difference between the expert score and all other groups confirms that there may be problems with the knowledge and attitudes of users. It appears likely that those with a bad attitude towards ratings will also know less about them. Ratings use is widespread and more than a third of our repondents indicated that they had participated in promotion and tenure process using ratings; a closer look at the responses of that subgroup indicated that between 25 and 50% of that group did not know the correct answer for most important knowledge items. For example, this subgroup typically did not realize what were the possible effects of student, course, and teacher variables, what the standard deviation represents, or the importance of using ratings that have been properly collected. It seems likely that this subgroup would frequently make incorrect decisions based on ratings.

The persistent association of attitude and knowledge does not appear to be related to teaching experience, rank, or institutional roles. However, attitude is not always explained by knowledge as in the case of School C which has the most positive attitude among the schools and the lowest knowledge results. School C's strong emphasis on its teaching mission as major community college system may help explain the low quantitative score for that group if faculty are less involved in research. Because School C has no institutional ratings system and is known for its support for teaching improvement, it is tempting to speculate that these respondents can afford positive attitudes since they have not been hurt by 'ratings malpractice' in promotion and tenure decisions and they have every reason to expect that ratings would be integrated into the general systems for teaching support.

For the "school" groups, of all the items contained in the knowledge subset, only the subscale dealing with criteria for collecting ratings data approached the score of the expert group.

The fact that no significant difference in knowledge (COMMON 2) was found between respondents who had participated in choosing, adapting, or writing ratings questionnaire forms versus those who hadn't, suggests that those who do participate may not refer to the literature or that their efforts to use the literature as a guide are not effective.

Some expert respondents, upon debriefing, indicated that in the "real world" less than perfect data must be used and therefore they did the best they could and did not choose the rigorous "uncertain" option for bad data. The practice subset results may be equivocal to some extent because the disagreements among experts are real. This will pose problems for practitioners who wish to obtain guidance from the literature.

Not unexpectedly, those in administrative roles have more positive attitudes toward rating others than those who only teach (and, consequently, get rated). For these groups, a positive climate for teaching seems to have gone hand in hand with a less negative attitude toward ratings.

Those reporting administrative and teaching duties score higher in knowledge than administrators with no teaching duties and teachers with no other duties. This is a source of concern since administrators who are likely to use ratings for merit review should be well enough informed to avoid using ratings incorrectly and faculty who submit them should know enough to defend their own interests. It is gratifying also that those who reported instructional and faculty development duties with or without teaching duties scored among the highest groups.

10. Conclusions:

Positive attitudes towards students as raters and the usefulness of ratings in general, were associated with higher knowledge scores. Since those respondents who have received assistance in interpreting ratings result also showed more positive attitudes and higher knowledge scores, it is encouraging to think that both a positive attitude toward and knowledge about ratings may be acquired directly through guided experience.

Taken together, the results from the institutional role vs. knowledge and the general positive relationship of knowledge with attitude for all subjects are also encouraging. They suggest that those who practice teaching consultation may know more about ratings than their clients and be of service to them, especially if those who receive consultation benefit from it through increased knowledge of ratings issues and positive attitudes toward ratings, as these results suggest they may. It is tempting to notice the general trend of higher knowledge and positive attitudes in the institutions with the most faculty development and instructional support services.

However, the magnitude of differences in overall knowledge scores suggest that many users of ratings are not sufficiently aware of the issues currently set forth in the ratings literature to be able to make decisions for summative or formative purposes. In particular, poor scores in subsets requiring knowledge of student, teacher, and course variables that may affect ratings outcomes, qualitative and quantitative knowledge needed for interpreting ratings "numbers", and the the evident negative attitudes of some users all add up to the need for closer examination of the users in action as they interpret and apply ratings.

Although evidence supporting the generalizability of our results is not offered, Schools A, B, and C at least represent three different environments for the assessment of teaching. The similarity of the their results and the distance between their results and those of the experts and the professional group underscores the need for further investigation.

This study offers no measure of the relative importance of the many concepts embodied in its surveys. Some concepts must be more critical for valid, reliable practice than others. Qualitative studies using direct observation and interviews may be helpful in understanding how users interpret ratings and apply them to decision-making. Large scale studies in institutions

that use ratings extensively could compare report designs as treatments so some measure of how well reports facilitated decision-making might be obtained. Perhaps debriefing expert interpreters can provide rules of thumb for users.

Unless personnel and teaching improvement processes incorporate ratings data in ways that insure decisions will not be made on the basis of chance differences in numbers, student ratings data are essentially meaningless. Ratings can only function as a sort of quantitative ink blot onto which users project what they already believe. In the context of teaching improvement consultation with a skilled consultant, decisions may depend less on ratings. Invalid interpretations may be ameliorated by classroom observation or other diagnostic strategies. However, for personnel decision-making where scarce resources frequently increase reliance on ratings data, real harm may be done to careers. In either case, scarce resources are squandered.

Persistent faculty paranoia about ratings and hostility toward their use may not be unfounded if those who use ratings are unqualified. In other words, those who reject ratings if they believe that ratings don't "work", may be justified but possibly for the wrong reasons. The issue is not the ability of students to provide valid, reliable information, but the ability of faculty and adminstrators to use it.

It may seem disingenuous to suggest that many people will not be surprised at these results: these data may confirm their worst suspicions about student ratings. Still, it is not the intention of this study to discredit student ratings. We began with a stake in one particular strategy for improving practice: finding ways for reports to communicate usable information to faculty, adminstrators, and teaching improvement specialists.

This study demonstrates that users do not always have knowledge and skills needed to use ratings, but it does not follow that they will have to obtain instruction or training. Because ratings exists in larger systems, we cannot reasonably expect every end user to be a statistician or have the psychometric skills to evaluate his/her own skill at interpreting ratings. Providing guidelines, warnings, interpretive statements, and comments may help users to understand numbers. To demonstrate that reports of results can compensate for missing skills will require further study. Whether or not improving reports of results can override other problems associated with poor practices is a question this study cannot address. We will assume that there are extrinsic factors that even the most knowledgable users will be unable to influence. For example, if institutional climate of support for teaching is poor, the stress of teaching may overwhelm any interest in assessing or improving it.

Evaluating the use of ratings could lead to the development of strategies to increase the competence of users and the ensuing likelihood that ratings will be used in a valid and reliable manner. Making ratings work may be essential for increased positive perception by faculty and administrators of the usefulness of ratings for teaching improvement and promotion and tenure decision-making. An increasingly positive attitudes toward the use of ratings combined with valid use of ratings data, could even establish the cycle of critical feedback and skilled practice that can produce what Levinson-Rose and Menges (1980) suggest is lacking in many teaching improvement activities: "...lasting changes in teacher behavior or lasting impact on students". (p.419)

References

Aleamoni, L.M. (1978) *The usefulness of student evaluations in improving college teaching.* Instructional Science, 7, (95-105).

Aleamoni, L.M. (1981) *Student ratings of instruction.* In J. Millman (ed.) *Handbook of teacher evaluation*, Beverly Hills, CA: Sage.

Bergquist, W.H. and Phillips, (1975) *Ahandbook for faculty development.* Washington, D.C.: Council for the Advancement of Small Colleges

Braskamp, L.A., Brandenburg, D.C., & Ory, J.C. (1984) *Evaluating teaching effectiveness.* Beverly Hills: Sage Publications.

Centra, J.A., Froh, R.C., Gray, P.J, Lambert, L.M., (1988) Diamond, R.M. (ed.) *A guide to evaluating teaching for tenure and promotion.* Syracuse, NY: Center for Instructional Development, Syracuse University.

Centra, J.A., (1980) *Determining faculty performance.* San Francisco, CA: Jossey BAss.

Cohen, P.A. (1981) *Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies.* Review of Educational Research. 51, (281-309).

Cranton, P.A. and Geis, G.L., (1982) *Evaluating teaching.* Montreal: Centre for Teacing and Learning Services, McGill University.

Doyle, K.O. Jr, (1975) *Student evaluation of instruction.* Lexington, MA: Lexington Books.

Doyle, K.O. Jr, (1983) *Evaluating teaching.* Lexington, Ma: Lexington Books.

Eble, K. E. (1986) *The craft of teaching:a guide to mastering the professor's art* (2nd ed.). San Francisco: Jossey-Bass Publishers,

Eble, K.E. (1985) *The aims of college teaching.* San Francisco:Jossey-Bass Publishers,

Furmann, Barbara S., and Grasha, Anthony F. (1983) *A Practical Handbook for College Teachers.* Boston: Little, Brown and Company.

Glass, G.V. and Hopkins, K.W. (1984) Statistical method in educational psychology. (2nd ed.). New Jersy: Prentice Hall Publishers.

Gullette, Margaret Morganroth, Ed. (1982) *The Art and Craft of Teaching.* Cambridge, MA: Harvard-Danforth Center for Teaching and Learning.

Levinson-Rose,J. & Menges, R. (1981) *Improving college teaching: a critical review of the research.* Review of Educational Research, 51, 403-434.

Lowman, Joseph. (1984) *Mastering the techniques of teaching.* San Francisco: Jossey Bass Publishers.

McKeachie, William J. (1978) *Teaching tips: a guidebook for the beginning college teacher.* 7th ed. Lexington, Ma:D.C. Heath and Company.

Miller, R.I., (1988) *Evaluating faculty for promotion and tenure,* San Francisco, CA: Jossey Bass.

Millman, J. (Ed), (1981). *Handbook of teacher evaluation.* beverly Hills, CA: Sage.

Seldin, P. (1980) *Successful faculty evaluation programs.* Crugers, N.Y.: Coventry Press.

Seldin, P. (1984) *Changing practices in faculty development.* San Francisco, CA: Jossey BAss.

Seldin, P. (1989) *How colleges evaluate professors.* Bulletin of the American Association of Higher Education. March 1989, 41, 7(3-7).

Appendix A: Sample Questionnaire (69 item form)

Appendix B:

Expert Researchers/writers/practitioners Who Participated in the First Validation of the Survey Item Bank

Dr Philip Abrami	Concordia University
Dr Lawrence Aleamoni	Arizona State University
Dr Raoul Arreola	University of Tennessee
Dr Dale Brandenberg	University of Illinois
Dr Mary Ann Bunda	Western Michigan University
Dr William Cashin	Kansas State university
Dr John Centra	Syracuse University
Dr Peter Cohen	Medical College of Georgia
Dr Patricia Cranton	Brock University
Dr Kenneth Doyle	University of Minnesota
Dr Glenn Erickson	University of Rhode Island
Dr Stanford Erickson	University of Florida
Dr Kenneth Feldman	SUNY Stonybrook
Dr Peter Frey	Northwestern University
Dr George Geis	Ontario Institute for Studies in Higher Education
Dr Gerry Gillmore	University of Washington
Dr Christopher Knapper	Waterloo University
Dr James Kulik	University of Michigan
Dr Wilbert McKeachie	University of Michigan
Dr Philip McKnight	University of Kansas
Dr Robert Menges	Northwestern University
Dr Harry Murray	University of Western Ontario
Dr Peter Seldin	Pace University
Dr Mary Deane Sorcinelli	Indiana University

SURVEY ON STUDENT RATINGS

INSTRUCTIONS: PART 1

Using the responses below, write the response that most closely matches your reaction to the statements in items 1 through 41 next to each item in the space provided.

A = strongly agree
B = tend to agree
C = tend to disagree
D = strongly disagree
E = uncertain, no opinion

1. There is rarely good agreement between the ratings of students and other measures of teaching competence (such as administrative reviews, peer observations, and interviews with students conducted by independent consultants).
2. Student rating questionnaires properly constructed and administered can do a good enough job of measuring teaching effectiveness to warrant their use for personnel decisions.
3. Student ratings can't really give a teacher the kind of feedback needed for improving the quality of instruction s/he provides.
4. Very specific items dealing with a teacher's behavior and skills usually correlate more highly with student achievement than do broad, global items such as "this instructor compared to others."
5. Really good, but "tough" teachers will tend to receive lower ratings and will only be appreciated by students years later when real life experiences have provided a new perspective.
6. Fifteen students in a given course are more likely to agree with each other in their rating of an instructor than are fifteen of the instructor's colleagues rating the same course.
7. Even a few written comments from students can often tell more about a course than a page full of statistics from a student rating questionnaire, no matter how many students responded to it.
8. Discussion, lecture, and laboratory classes are so different that there is no valid way to use student ratings to compare the instructors who teach them.
9. Student ratings used for personnel purposes should not be returned to the instructor until after final grades have been submitted.
10. When student ratings are used for promotion and tenure decision-making, it really doesn't matter who administers the course evaluation.
11. Students should be required to sign their answer sheets.
12. The instructor should leave the room during the administration of a student rating questionnaire.
13. When ratings are mandatory, every course/section of every instructor should be evaluated every academic term.
14. Telling students that results will be used for personnel purposes tends to cause students to be more critical and thus lowers ratings.
15. There is no predictable relationship between student ratings and student achievement.
16. Students are more likely to give slightly higher ratings to their majors or electives than to courses taken to fulfill a college requirement
17. The ratings obtained from "good" students should be given more weight than those of students who aren't doing well, i.e. the higher the grade point average of the rater, the more likely valid that rater's responses.
18. A student's age and sex tend to exert predictable and systematic influences on the ratings s/he gives.
19. There are no predictable differences between the ratings given to lower level (beginning) courses and those given to upper level courses.
20. Students who expect higher course grades tend to give more positive ratings than students who expect lower course grades.
21. Charismatic, entertaining teachers tend to get better ratings on "overall" items (e.g. "this instructor compared to others" and "this course compared to others") than less "flashy," but otherwise better teachers.
22. An otherwise poor teacher can get higher ratings by lenient grading.
23. Generally, a rating of heavier than average workload and/or difficulty is associated with lower than average ratings of the teacher's performance based on "overall" item scores.
24. The content area of courses is associated with predictable differences in the ratings students give.
25. Class size has no predictable influence on student ratings.
26. Adminstrators should usually give more weight to student rating items dealing with an instructor's specific teaching skills (such as communication, rapport, testing, and grading) than to broad, global items such as "this instructor's teaching effectiveness compared to others."

27. Properly constructed, administered, and analyzed student ratings are usually sufficient as the sole source of information about teaching effectiveness for the purpose of personnel decision-making.

28. Assuming that an adequate number of students respond to have a reasonably representative sample, the smaller the classes, the more classes are needed to determine an instructor's "average" teaching performance.

29. The number of courses needed to provide a reasonable basis for understanding a teacher's performance depends largely on how the rating results are to be used.

30. Given a valid questionnaire properly administered, the ratings obtained in a single large lecture course (e.g. 100 students) would be an adequate sample of student opinion for use in personnel decision-making, provided that at least 75% of the students responded.

31. The proportion of a class that rates an instructor is not as important as the total number of raters.

32. The best measure of overall teaching performance is obtained by averaging the scores for each item on the questionnaire into one 'grand mean'.

33. When interpreting ratings results, one way to help control the influence of bias in student ratings due to factors beyond the instructor's control is to consider scores for any available, relevant comparison groups.

34. For personnel decision-making, nationally based norms are a fairer basis for comparison than local (institutional or departmental) norms.

35. A good way to understand both the range and direction(s) of student opinion in a class is to examine the percentages of students responding to each option for an item.

36. For items with a five point scale, standard deviations of less than 1.5 generally indicate that the respondents were in relatively good agreement on an item.

37. For items with a five point scale, standard deviations of more than 1.7 make the results for that item nearly useless for interpretation for any purpose.

38. For student ratings, a mean score with a percentile ranking of 65 usually indicates better teaching performance than a mean with a percentile ranking of 50.

39. For an item such as "rate this instructor compared to others you have had," a mean score with a percentile ranking of 10 indicates a poor teaching performance.

40. Student ratings usually fall within a classic "bell curve" so that the "average" rating on an item with a response scale of 1 to 5 pts. would be very close to 3.00.

41. When comparing an instructor's ratings to those of other instructors, standardized scores such as z-scores or t-scores are a better measure than directly comparing item means.

INSTRUCTIONS: PART 2

For items 42 - 51, you are a member of department XYZ's promotion and tenure committee reviewing the teaching effectiveness of five instructors. Many sources of evidence concerning their teaching performance have already been considered by the committee. The committee's attention now turns to the available student ratings for the five. The ratings were obtained using a short questionnaire which included the three items below. The results are summarized in the table on the facing page.

KEY TO QUESTIONNAIRE ITEMS:

STUDENT'S SELF-REPORT OF AMOUNT LEARNED

5 pts	exceptional amount
4 pts	more than usual
3 pts	about as much as usual
2 pts	less than usual
1 pt	almost nothing

STUDENT'S OVERALL RATING OF INSTRUCTOR

5 pts	among the best
4 pts	better than average
3 pts	about average
2 pts	worse than average
1 pt	among the worst

STUDENT'S OVERALL RATING OF COURSE

5 pts	among the best
4 pts	better than average
3 pts	about average
2 pts	worse than average
1 pt	among the worst

At various times during the course of deliberations, the statements at the bottom of the facing page are made by members of the committee. Based on the data in this report alone, use the response options below to indicate your agreement with each statement (i.e. items 42 to 51).

A = strongly agree
B = tend to agree
C = tend to disagree
D = strongly disagree
E = uncertain, or can't determine based on the available ratings data.

REPORT OF STUDENT RATINGS OF INSTRUCTION IN THE XYZ DEPARTMENT

The average ratings in the XYZ department (308 courses in sample) are:	Mean*	S.D.	Median	High - Low**	The average ratings in the University (4069 courses in sample) are:	Mean	S.D.	Median	High - Low
LEARNED:	3.53	0.5	3.58	4.80 - 2.26	LEARNED:	3.63	0.5	3.69	5.00 - 1.33
INSTRUCTOR:	3.60	0.7	3.67	5.00 - 1.50	INSTRUCTOR:	3.75	0.7	3.86	5.00 - 1.38
COURSE:	3.47	0.5	3.53	4.68 - 1.89	COURSE:	3.62	0.6	3.67	5.00 - 1.56

* mean of section means
** highest and lowest section mean in sample

PROFESSOR SMITH:

COURSE:	LEARNED Mean	S.D.	T Rank	INSTRUCTOR Mean	S.D.	T Rank	COURSE Mean	S.D.	T Rank	enrolled/responded
XYZ 101-01	3.78	0.7	55 HI	4.45	0.7	62 HH	3.63	0.7	53 HI	43/40 (93%)
XYZ 101-03	3.43	0.7	42 HH	4.71	0.5	66 HI	3.32	1.1	45 HI	30/28 (93%)
XYZ 101-02	3.29	0.7	43 LH	4.29	0.9	49 HI	3.41	0.9	49 HI	42/34 (81%)

PROFESSOR ASPEN:

COURSE:	LEARNED Mean	S.D.	T Rank	INSTRUCTOR Mean	S.D.	T Rank	COURSE Mean	S.D.	T Rank	enrolled/responded
XYZ 270-01	3.67	0.5	53 HI	4.50	0.5	63 HH	4.17	0.4	64 HI	7/6 (86%)
XYZ 402-01	4.20	0.7	65 HI	4.80	0.4	67 HI	3.60	0.5	53 HI	5/5 (100%)
XYZ 401-01	4.17	0.7	64 HI	4.33	0.7	60 HH	4.17	0.7	64 HI	7/6 (86%)
XYZ 401-02	4.20	0.7	65 HI	4.60	0.8	64 HI	4.29	0.7	66 HI	5/5 (100%)
XYZ 402-01	4.80	0.4	78 HI	5.00	0.0	70 HI	4.60	0.5	70 HI	5/5 (100%)

PROFESSOR FAHEY:

COURSE:	LEARNED Mean	S.D.	T Rank	INSTRUCTOR Mean	S.D.	T Rank	COURSE Mean	S.D.	T Rank	enrolled/responded
XYZ 401-01	3.30	1.0	47 HI	3.30	1.0	46 HI	3.61	0.8	53 HI	53/23 (43%)
XYZ 401-01	3.50	0.9	49 HI	3.50	0.8	49 HI	3.36	0.9	48 HI	35/18 (51%)
XYZ 401-01	3.35	1.1	46 HI	2.76	1.1	38 LH	2.85	0.9	38 LH	61/29 (48%)
XYZ 401-01	3.97	1.4	59 HH	3.57	0.9	50 HI	3.63	0.7	53 HI	39/16 (41%)

PROFESSOR COHEN:

COURSE:	LEARNED Mean	S.D.	T Rank	INSTRUCTOR Mean	S.D.	T Rank	COURSE Mean	S.D.	T Rank	enrolled/responded
XYZ 402-01	3.36	1.0	45 LH	3.00	1.0	42 LH	3.48	1.0	50 HI	29/23 (79%)
XYZ 402-01	3.83	0.8	57 HH	3.67	1.0	51 HI	3.50	0.6	51 HI	14/12 (86%)
XYZ 402-01	3.57	0.8	51 HI	3.43	1.1	48 HI	3.70	0.5	55 HI	30/23 (77%)
XYZ 402-01	4.33	0.6	67 HI	4.08	0.8	57 HH	4.17	0.7	64 HI	16/12 (75%)
XYZ 402-01	3.55	0.7	50 HI	3.60	0.8	50 HI	3.75	0.8	51 LO	29/21 (72%)

PROFESSOR PARKER:

COURSE:	LEARNED Mean	S.D.	T Rank	INSTRUCTOR Mean	S.D.	T Rank	COURSE Mean	S.D.	T Rank	enrolled/responded
XYZ 402-01	3.49	0.8	49 HI	4.32	0.8	60 HH	3.65	0.8	54 HI	38/37 (95%)
XYZ 402-01	3.82	0.8	56 HH	4.00	0.9	57 HH	3.86	1.0	58 HH	33/22 (67%)

42. Aspen is more effective than Smith.
43. Smith is more effective than Fahey.
44. Cohen is less effective than Fahey.
45. Cohen is less effective than Aspen.
46. Parker is more effective than Fahey or Cohen.
47. Fahey is the least effective of the group.
48. If only one of these instructors were to receive a merit raise, Aspen is the best choice of the five.
49. It is not fair to compare Smith and Cohen with each other.
50. Instructors in XYZ department are rated relatively lower than instructors in the university at large.
51. With the possible exception of Smith and Cohen, it is impossible to fairly compare the overall teaching effectiveness of any two instructors within this group, based on the ratings presented in this report.

INSTRUCTIONS: PART 3

Use the response scale below for the items 52 - 63:

A = Strongly agree
B = Agree
C = Slightly agree
D = Slightly disagree
E = Disagree
F = Strongly disagree

52. Student rating questionnaires are not a very useful way to measure teaching effectiveness in my particular discipline.

53. I know enough about statistics to interpret the results of student ratings without assistance.

54. I don't really know how to use the results of student ratings to improve my teaching skills or course.

55. Administrators generally use student ratings in promotion and tenure decisions in ways that accurately depict quality of teaching performance.

56. Students have a right/responsibility to rate courses/instructors.

57. The use of student ratings in my institution provides little or no benefit to the quality of instruction students receive.

58. I believe mandatory evaluation of teaching performance without institutional support for teaching improvement is punitive.

59. I trust my students to give sincere and honest responses on student rating questionnaires.

60. I have been able to improve my teaching or my course based on information I obtained from student ratings.

61. With the exception of notably poor teaching performance, there appears to be little or no practical administrative interest in the quality of teaching in my department.

62. I am not aware of any explicit procedures or policies regarding the use of student ratings in personnel decisions in my department.

63. I feel my career has been harmed to some degree by student ratings I have received.

64. Indicate the response which best fits your current duties (including research).

A = teaching only
B = teaching with administrative duties, e.g. department chair
C = primarily administrative appointment with some teaching duties
D = administrative appointment, no teaching duties
E = instructional development, faculty development, or evaluation professional with some teaching duties
F = instructional development, faculty development, or evaluation professional, no teaching duties

65. How many years have you been teaching at the post-secondary level?

A = less than one
B = one to two
C = more than two but less than eight
D = eight or more but less than twelve
E = more than twelve but less than twenty
F = twenty or more

66. Indicate your faculty rank. (Leave blank if non-faculty appointment.)

A = full professor (including emeritus)
B = associate professor
C = assistant professor
D = instructor
E = lecturer (including adjunct, senior, and part-time)
F = teaching assistant

67. Indicate how often you have participated in personnel decision making in which student ratings (other than your own) were offered as evidence of teaching performance.

A = on a regular basis (at least once every academic year for at least the last two years)
B = not currently, but have done so at least twice in my career
C = have never done so

68. Indicate which response describes how you have participated in creating or selecting student rating forms

A = I have helped choose forms for use in my department or institution for personnel decision making.
B = I have helped write or adapt forms for use in my department or institution for personnel decision making.
C = I have chosen, adapted, or written forms for use by others in my department or institution for diagnosis for teaching improvement.
D = I have chosen or written forms for my own use to present evidence of my teaching performance for personnel decision-making.
E = I have chosen or written forms for my own use in obtaining diagnostic feedback for teaching improvement.
F = I have never selected, adapted, or written such forms.

69. Have you ever received assistance from a teaching improvement specialist or master teacher in your efforts to interpret and/or use student ratings to improve the instruction you provide?

A = YES, on at least two separate occasions
B = YES, but only once
C = NO, never, BUT I would like to try it sometime
D = NO, never, AND I am not interested in trying it